Utilisation des statistiques pour les sciences de la vie

Gaëlle LELANDAIS

gaelle.lelandais@universite-paris-saclay.fr

Ce document est mis à disposition selon les termes de la licence Creative Commons Attribution – Partage dans les mêmes conditions 4.0 International (CC BY-SA 4.0).

Utilisation des statistiques pour les sciences de la vie

Table des matières

Partie 1 : Pourquoi utiliser les statistiques pour les sciences de la vie ?

I. Présentation des cas d'études

1) Contexte de travail dans un laboratoire pharmaceutique : « le médicament est-il efficace ? »

Une entreprise pharmaceutique développe un médicament destiné à traiter l'hypertension artérielle. Le protocole expérimental consiste à mesurer la tension artérielle systolique (en cm Hg) deux fois chez des patient(e)s : une première fois avant le traitement (valeur de référence) et une deuxième fois après 6 semaines de prise du médicament.

- L'objectif est de mettre en place un protocole d'étude statistique destiné à évaluer l'efficacité du nouveau médicament.

2) Contexte de travail dans un casino[1] : « les jetons sont-ils équilibrés ? »

Un casino propose le jeu « pile/face » à ses client(e)s. Des centaines de jetons sont livrés au casino et la personne qui dirige l'établissement souhaite vérifier qu'ils sont équilibrés, soit : $\Pr(pile) = \Pr(face)$[2]. Le casino possède une machine qui lance des jetons et détecte automatiquement le côté gagnant.

- L'objectif est de mettre en place un protocole d'étude statistique destiné à contrôler les nouveaux jetons. En particulier, il faut déterminer 1) combien de fois lancer un jeton et 2) combien de jeton lancer.

[1] Ce deuxième contexte de travail n'est pas en relation direct avec les sciences de la vie. Toutefois, il est utile pour discuter le cas particulier de l'étude des variables qualitatives à deux classes, rencontrées fréquemment en biologie.

[2] La probabilité d'obtenir le côté « Pile » est égale à la probabilité d'obtenir le côté « Face », et égale à ½.

3) Premiers résultats

Une personne malade prend le médicament développé par l'entreprise pendant 6 semaines. Sa tension artérielle[1] est de 15 cm Hg avant le traitement et 12 cm Hg après le traitement (soit une diminution de 3 cm Hg).

➢ **Quelle conclusion est-il possible d'obtenir à partir de ce résultat ? Discuter la notion de REPRODUCTIBILITE en biologie.**

9 personnes supplémentaires prennent le médicament pendant 6 semaines. Pour chacune, les valeurs de la tension artérielle sont mesurées avant et après le traitement :

Personne	1	2	3	4	5	6	7	8	9	10
Avant	15	15	16	18	18	18	18	15	19	18
Après	12	13	16	16	14	15	17	14	16	16
Écart	3	2	0	2	4	3	1	1	3	2

Tableau 1 : Mesures de la tension artérielle (en cm Hg) de 10 personnes malades, avant et après la prise du médicament. Les écarts correspondent à la différence : tension artérielle « avant » moins tension artérielle « après » le traitement. La valeur moyenne des écarts est de 2,1 et la variance 1,29 (soit un écart type de 1,13). Des valeurs d'écarts positives sont attendues, si le traitement est efficace.

➢ **Quelle conclusion est-il possible d'obtenir à partir de ces résultats ?**

4) Bilan

- **Ce qui est certain** : des différences sont observées entre les mesures de la tension artérielle après et avant traitement.

- **Ce qui est hypothétique** : l'origine de ces différences (efficacité du médicament ou bien hasard d'échantillonnage ?).

[1] Notez que les valeurs données tout au long de ce polycopié ont été choisies pour illustrer des concepts statistiques. Leur pertinence d'un point de vue médical peut-être contestable ☺.

II. Problématique statistique

A partir d'un ensemble d'observations, l'objectif est de déterminer :

- la part (obligatoire) du hasard,
- la part (éventuelle) de l'effet testé (efficacité du médicament par exemple).

○ Les statistiques permettent de calculer la probabilité d'observer uniquement par le hasard des écarts au moins aussi importants que ceux observés lors de l'étude. Ainsi : **Statistiques → Calculer la probabilité du hasard.**

Le raisonnement est alors le suivant :

➢ Si la probabilité d'observer uniquement par le hasard les données du Tableau 1 est grande, l'effet du médicament a peu de chances d'exister. En revanche,

➢ Si la probabilité d'observer uniquement par le hasard les données du Tableau 1 est petite, l'effet du médicament a de grandes chances d'exister[1].

➢ **Quels paramètres influencent le résultat d'une étude statistique ?**

III. Rappels de définitions

1) Population et unités statistiques

La population est l'ensemble des éléments qui forment le champ d'analyse d'une étude particulière. Elle est constituée d'un ensemble d'éléments appelés unités statistiques[2] (ou individus).

➢ **Quelles sont les populations étudiées dans les cas d'études ?**

[1] Cette notion est difficile à appréhender. Elle devra être reprise à l'issue du CHAPITRE « Tests d'hypothèses » (page 31).

[2] Le concept de population est général en statistiques. Il s'applique aux choses, aux événements (pas seulement aux êtres humains).

2) Variables

Les unités statistiques d'une population varient selon des caractéristiques appelées des variables. Les variables peuvent être quantitatives (discrètes/continus) ou qualitatives (2 classes ou plus).

➢ **Quelles sont les variables étudiées dans les cas d'études ? Sont-elles quantitatives ou qualitatives ?**

(a) Présentation des mesures d'une variable quantitative

- **Moyenne**

La moyenne d'une variable quantitative X est :

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Équation 1 : Formule de la moyenne d'une variable quantitative.

- **Variance**

La variance d'une variable quantitative X est :

$$Var = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Équation 2 : Formule de la variance d'une variable quantitative.

➢ **Calculer les valeurs de la moyenne et de la variance associées aux données du Tableau 1.**

➢ **Que signifie une valeur faible de variance en termes de reproductibilité des mesures ?**

- **Représentation histogramme**

Un ensemble de mesures d'une variable peut être représenté sous la forme d'un histogramme, avec en abscisse des intervalles de valeurs de la variable étudiée et en ordonnée les nombres de mesures (effectifs) observés dans ces intervalles.

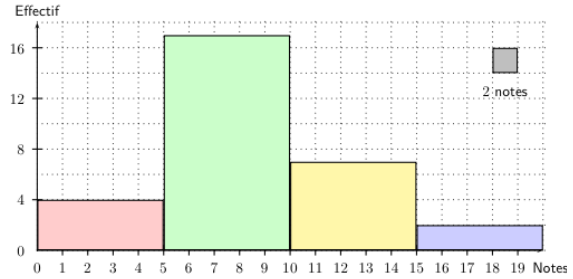

Figure 1 : Représentation histogramme du variable quantitative (« les notes des étudiants à l'examen »). Cette image est issue de http://mathematiques.daval.free.fr/IMG/pdf/Statprobas_2_series_continues.pdf.

IV. Expérience et variable aléatoire

1) Définitions

Une expérience aléatoire est une expérience qui, répétée dans des conditions apparemment identiques, peut donner des résultats différents (nommés événements).

➢ **Quelles sont les expériences aléatoires réalisées dans les cas d'études ?**

Une variable aléatoire est réalisée lorsqu'elle prend des valeurs en fonction du résultat d'une expérience aléatoire.

➢ **Quelles sont les variables aléatoires[1] associées aux expériences aléatoires des cas d'études ?**

[1] Décrire une variable aléatoire signifie donner d'une part son intitulé et d'autre part son type.

2) Convention de notations

Une variable aléatoire est notée en MAJUSCULES, tandis que les valeurs prises lors de la réalisation de cette variable aléatoire (c'est à dire les observations) sont notées en minuscules. Ainsi, les données du

Personne	1	2	3	4	5	6	7	8	9	10
Avant	15	15	16	18	18	18	18	15	19	18
Après	12	13	16	16	14	15	17	14	16	16
Écart	3	2	0	2	4	3	1	1	3	2

Tableau 1 (page 7) sont notées :

- X : variable aléatoire « écart en les mesures de la tension artérielle avant et après le traitement »,
- Et $\{x_1 = 3; x_2 = 2; x_3 = 0; x_4 = 2; x_5 = 4; x_6 = 3; x_7 = 1; x_8 = 1; x_9 = 3; x_{10} = 2\}$ les mesures de X dans l'échantillon (ici de taille 10).

V. Notion d'échantillonnage

Une étude statistique a pour objectif d'obtenir des connaissances sur l'ensemble de la population (le médicament sera-t-il efficace sur toutes les personnes malades ?). Une analyse complète de la population est souvent impossible (la population totale est inconnue, mal définie, ou bien elle comprend trop d'individus).

Un échantillonnage consiste à choisir parmi les éléments de la population un ensemble d'unités statistiques[1] pour lesquels des observations sont réalisées. Les unités statistiques choisies constituent un échantillon. Si l'échantillon est bien choisi, les observations permettent d'obtenir des connaissances sur la population.

➢ **Comment constituer un échantillon informatif sur le plan statistique ?**

[1] La définition d'une unité statistique est donnée page 11.

Partie 2 : Calculs de probabilités en statistiques

I. Pourquoi des probabilités en statistiques ?

La valeur prise par une variable aléatoire à l'issue d'une expérience aléatoire, ne peut pas être « prédite » avec certitude. Toutefois, il est possible de calculer la probabilité des différentes valeurs possibles. Également (voir page 8) :

- Les statistiques permettent de calculer la probabilité d'observer uniquement par le hasard des écarts au moins aussi importants que ceux observés lors de l'étude. Ainsi : **Statistiques → Calculer la probabilité du hasard.**

II. Variables aléatoires quantitatives discrètes[1]

1) Loi de probabilité

La loi de probabilité d'une variable aléatoire quantitative discrète est la fonction qui associe à chaque valeur x de la variable aléatoire X sa probabilité :

$$p(x) = P(X = x)$$

Équation 3 : Définition d'une loi de probabilité.

Les probabilités pour l'ensemble des valeurs x de X vérifient :

$$0 \le p_i \le 1 \text{ et } \sum_{i=1}^{n} p_i = 1$$

➢ **La loi de Bernoulli et de la loi Binomiale sont des exemples de lois de probabilité. Les connaissez-vous ?**

[1] Les résultats présentés dans ce paragraphe sont également applicables aux variables aléatoires qualitatives.

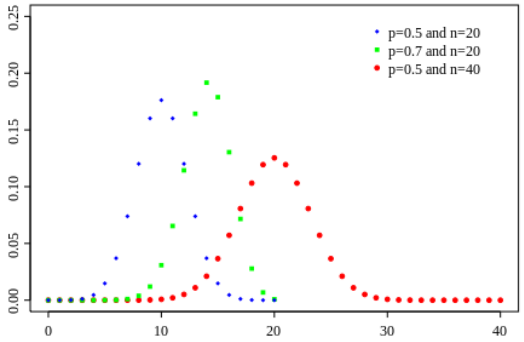

Figure 2 : Exemples de lois binomiales de variables aléatoires quantitatives discrètes. Ces lois binomiales sont définies par différentes valeurs de paramètres p et n. Les valeurs des variables aléatoires X sont en abscisse et les valeurs P(X = x) sont en ordonnée. Cette image est issue de http://fr.wikipedia.org/wiki/Loi_binomiale .

2) Fonction de répartition

La fonction de répartition d'une variable aléatoire X quantitative discrète est la fonction qui donne la probabilité que X soit inférieure ou égale à une valeur x donnée :

$$F(x) = P(X \leq x)$$

Équation 4 : Définition d'une fonction de répartition. La variable aléatoire X est quantitative.

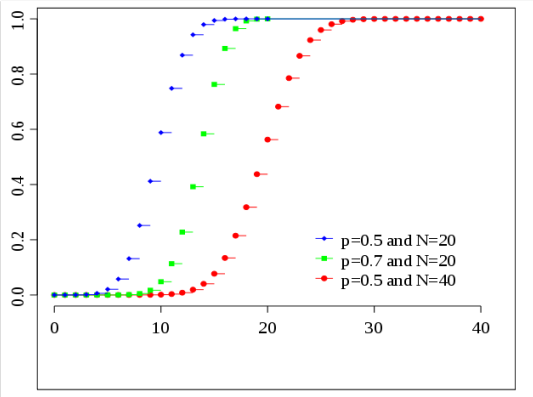

Figure 3 : Fonctions de répartition des variables aléatoires qui suivent les lois binomiales présentées Figure 2 (ci-dessus). Cette image est issue de http://fr.wikipedia.org/wiki/Loi_binomiale.

3) Espérance et variance

(a) Définitions

L'espérance d'une variable aléatoire X (discrète) est égale à la moyenne des valeurs de pondérées par leurs probabilités d'apparition :

$$E(X) = \sum_{i=1}^{n} x_i \times P(X = x_i)$$

Équation 5 : Formule de l'espérance. La variable aléatoire X est discrète.

➢ **Quelle est la différence entre « espérance » et « moyenne » ?**

La variance d'une variable aléatoire X (discrète) est égale à la somme des écarts à l'espérance, au carré et multipliée par les probabilités d'apparitions :

$$Var(X) = \sum_{i=1}^{n} (x_i - E(X))^2 \times P(X = x_i) = E\left(X^2\right) - \left(E(X)\right)^2$$

Équation 6 : Formules de la variance. La variable aléatoire X est discrète.

➢ **Pourquoi la racine carrée de la variance, appelée écart type, est souvent utilisée ?**

(b) Quelques propriétés

Soient a et b deux constantes et X une variable aléatoire. On a :

$$E(aX + b) = aE(X) + b$$

Équation 7 : Propriété de l'espérance d'une variable aléatoire.

Soient X et Y deux variables aléatoires. L'espérance mathématique d'une somme est égale à la somme des espérances mathématiques. Tandis que l'espérance mathématique d'une différence est égale à la différence des espérances mathématiques :

$$E(X+Y)=E(X)+E(Y)$$
$$E(X-Y)=E(X)-E(Y)$$

Équation 8 : Propriétés de l'espérance de deux variables aléatoires.

Soient X et Y deux variables aléatoires indépendantes[1] :

$$E(X.Y)=E(X).E(Y)$$

Équation 9 : Propriétés de l'espérance de deux variables aléatoires indépendantes.

Soient a et b deux constantes et X une variable aléatoire :

$$Var(aX+b)=a^2Var(X)$$

Équation 10 : Propriété de la variance d'une variable aléatoire.

Soient X et Y deux variables aléatoires indépendantes. La variance de leur somme est égale à la somme des variances. De même pour la différence :

$$Var(X+Y)=Var(X)+Var(Y)$$
$$Var(X-Y)=Var(X)+Var(Y)$$

Équation 11 : Propriétés de la variance de deux variables aléatoires indépendantes.

➢ **Quelles significations peut-on donner à ces propriétés en termes de reproductibilité des résultats d'une étude statistique ?**

[1] "L'indépendance est une notion probabiliste qualifiant de manière intuitive des événements aléatoires n'ayant aucune influence l'un sur l'autre" (Wikipedia, page "Indépendance (probabilités)).

III. Variables aléatoires quantitatives continues

1) Loi(s) normale(s)

(a) Représentation graphique

La loi normale s'applique à des variables aléatoires quantitatives continues définies entre $-\infty$ et $+\infty$. Elle est entièrement décrite par deux paramètres : la moyenne μ et la variance σ^2.

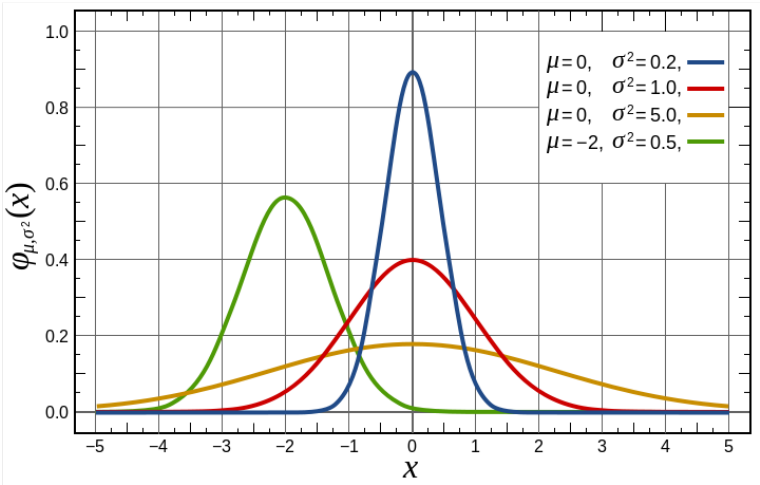

Figure 4 : Exemples de lois normales de variables aléatoires quantitatives continues. Ces lois normales sont définies par différentes valeurs de paramètres μ et σ^2. Les valeurs des variables aléatoires sont en abscisse et les valeurs $f(x)$ sont en ordonnée. Cette image est issue de http://fr.wikipedia.org/wiki/Loi_normale .

(b) Fonction de densité de probabilité

La densité de probabilité f d'une variable aléatoire quantitative continue est la probabilité calculée sur une variation faible (dx) de la variable aléatoire X :

$$P(x \le X \le x+dx) = f(x)dx$$

Équation 12 : Définition d'une fonction de densité de probabilité.

Dans le cas de la loi normale, la fonction de densité de probabilité s'écrit[1] :

$$f(x) = \frac{1}{\sqrt{\sigma^2 2\Pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

Équation 13 : Fonction de densité de probabilité de la loi normale.

(c) Notation

Une variable aléatoire quantitative continue qui suit une loi normale, par exemple de moyenne 12 et de variance 10 est notée : $X \sim \mathrm{N}(\mu = 12, \sigma^2 = 10)$. Sa fonction de densité de probabilité est représentée ci-dessous par la courbe noire. La courbe verte correspond à la fonction de densité de probabilité d'une variable aléatoire de même moyenne (= 12) mais de variance plus élevée (= 20). A noter que puisque la loi normale est une fonction de densité de probabilité :

- Les valeurs de $f(x)$ sont toutes positives
- L'aire totale sous la courbe est égale à 1.

Également, si X est une variable quantitative continue, alors $P(X = x) = 0$.

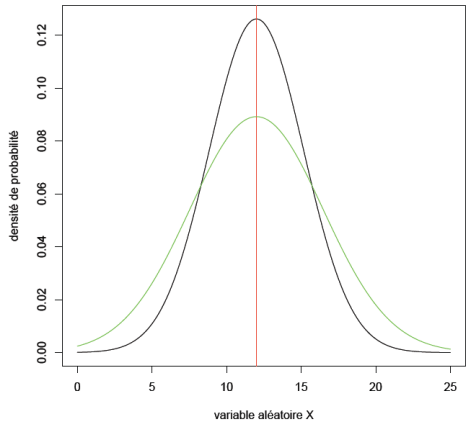

Figure 5 : Représentation de la loi normale de moyenne 12 et de variance 10 (en noire) et de la loi normale de moyenne 12 et de variance 20 (en vert). Une ligne verticale en $x = 12$ est représentée en rouge. Dans les deux cas les aires sous les courbes sont égales à 1.

[1] Cette formule est donnée seulement à titre indicatif. Elle est très peu utilisée dans la pratique.

(d) Fonction de répartition, fonction de densité et calculs de probabilités

La fonction de répartition $F(x)$ est une primitive de la densité de probabilité $f(x)$. La probabilité que X prenne une valeur comprise entre les bornes a et b est donc égale à :

$$P(a \le X \le b) = \int_a^b f(x)dx = F(b) - F(a)$$

Équation 14 : Relation entre fonction de répartition et fonction de densité de probabilité en termes de calcul de probabilité.

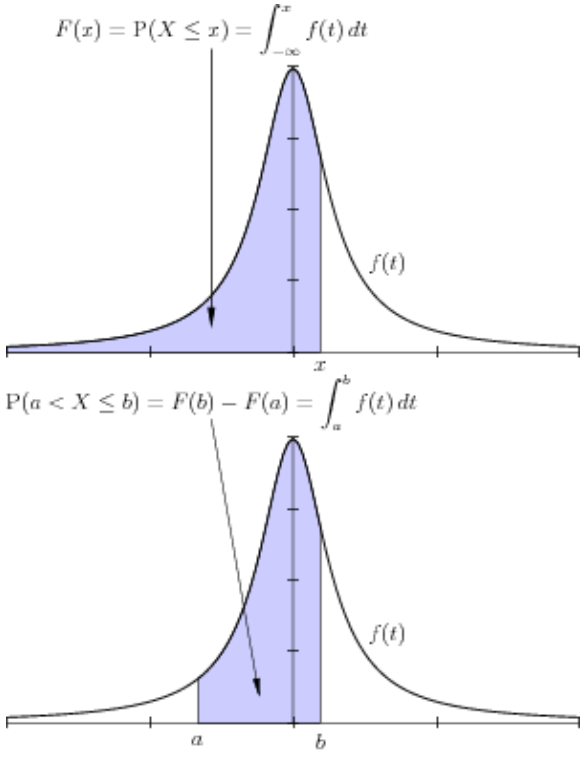

Figure 6 : Surface représentant une probabilité. Sur cet exemple la surface hachurée correspond à la probabilité que X prenne une valeur entre a et b. A noter qu'il est indifférent d'inclure ou d'exclure les bornes dans le calcul de probabilité d'un intervalle, lorsque la fonction de probabilité est continue. En effet $P(X = x) = 0$ si X est continue. Cette image est issue de http://fr.wikipedia.org/wiki/Densit%C3%A9_de_probabilit%C3%A9 .

2) Loi normale centrée – réduite

(a) Fonction de densité de probabilité

La loi normale centrée - réduite est une loi normale particulière de moyenne 0 et de variance 1 (voir ci-dessous). Les variables aléatoires qui suivent une loi normale centrée réduite sont souvent notées Z.

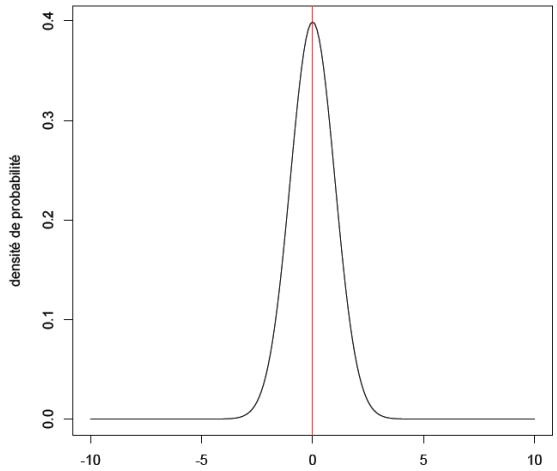

Figure 7 : Représentation de la loi normale centrée réduite. Les variables aléatoires qui suivent une loi normale centrée réduite sont souvent notées Z.

(b) Fonction de répartition

La fonction de répartition permet de calculer la probabilité que la variable aléatoire Z ait une valeur inférieure ou égale à une quantité quelconque z (voir page 13 et Figure 8 ci-dessous). Cette fonction de répartition est souvent notée ϕ :

$$\phi(z) = P(Z \leq z) \text{ avec } Z \sim N(0,1)$$

Équation 15 : Fonction de répartition de la loi normale centrée réduite.

➢ **Pouvez-vous vérifier graphiquement la relation suivante : $\phi(-z) = 1 - \phi(z)$.**

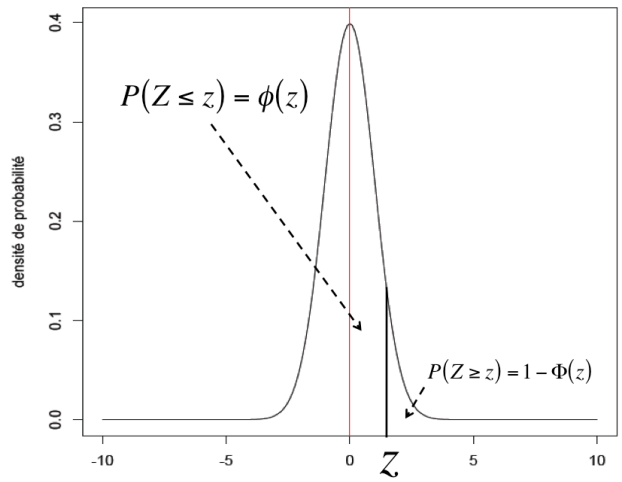

Figure 8 : Illustration d'un calcul de probabilité en utilisant la fonction de répartition ϕ de la loi normale centrée réduite.

3) Calculs de probabilités avec la loi normale

(a) Lecture de la table de la loi normale

Les valeurs de la fonction de répartition de la loi normale centrée réduite sont disponibles dans des tables statistiques, pour les valeurs positives de :

z	0.00	0.01	0.02	0.03	0.04
0.0	.5000	.5040	.5080	.5120	.5160
0.1	.5398	.5438	.5478	.5517	.5557
0.2	.5793	.5832	.5871	.5910	.5948
0.3	.6179	.6217	.6255	.6293	.6331
0.4	.6554	.6591	.6628	.6664	.6700
0.5	.6915	.6950	.6985	.7019	.7054
0.6	.7257	.7291	.7324	.7357	.7389

Figure 9 : Extrait de la table de la loi normale centrée-réduite (ou table de Gauss). En ligne et en colonne, les valeurs de Z sont notées. Les valeurs correspondantes $\Phi(z)$ sont dans le tableau. Ainsi si $z = 0,32$ alors $\Phi(z) = 0,6255$. Le reste de la table est présenté en Annexe.

(b) Procédure de « centrage/réduction » d'une variable aléatoire

Une variable aléatoire X qui suit une loi normale quelconque ($X \sim N(\mu,\sigma^2)$) peut être convertie en variable aléatoire Z qui suit une loi normale centrée réduite ($Z \sim N(0,1)$) en effectuant le changement de variable :

$$Z = \frac{X-\mu}{\sigma}$$

Équation 16 : Changement de variable associé à la procédure de « centrage / réduction » d'une variable aléatoire qui suit une loi normale quelconque.

La procédure de centrage – réduction permet de calculer des probabilités associées à une loi normale quelconque, pour diverses valeurs de μ et de σ^2 (il est toujours possible de se rapporter à une loi normale centrée réduite).

- **Discuter la notion de « Z-Score » couramment utilisée en biologie.**

(c) Applications

- **Soit X une variable aléatoire qui suit une loi normale de moyenne $\mu = 4$ et la variance $\sigma^2 = 4$, calculer $P(X \leq 6)$?**
- **Calculer $P(1 \leq X \leq 3)$ si $X \sim N(\mu = 2, \sigma^2 = 4)$. A noter que pour ces calculs il est possible d'utiliser des logiciels statistiques tel que R[1].**

IV. Théorème central limite

Le théorème central limite établit le lien entre la loi normale et une grande classe de lois de probabilité, quand le nombre d'observations augmente. C'est l'un des théorèmes les plus importants en statistiques.

[1] http://cran.r-project.org/

Énoncé du théorème :

Toute somme de n variables aléatoires indépendantes et qui suivent des lois de probabilité (ou des fonctions de probabilités) identiques, tend vers une variable aléatoire dont la loi converge vers la loi normale quand n augmente.

- **Pourquoi la plupart des variables étudiées en biologie suivent une loi normale ?**

Partie 3 : Échantillonnage et estimation

I. Définition statistique de l'échantillon

1) Étude d'une variable aléatoire quantitative[1]

Soit X une variable aléatoire de moyenne μ et de variance σ^2 (dans la population). La création d'échantillon de taille n est la réalisation indépendante, n fois, de la variable aléatoire X. C'est une expérience aléatoire. Un échantillon est donc une variable aléatoire à n dimensions, notée : $(X_1, X_2, X_3, ..., X_n)$.

➢ **Illustrer cette notion par un schéma. Représenter d'une part la population et d'autre part les différents échantillons possibles, issus de cette population. Respecter les conventions de notations (les variables aléatoires sont notées en MAJUSCULES, tandis que les valeurs prises par une variable aléatoire, les observations, sont notées en minuscules).**

2) Moyenne d'un échantillon

La moyenne des valeurs de X dans l'échantillon est également une variable aléatoire quantitative, notée :

$$M = \frac{X_1 + X_2 + ... + X_n}{n}$$

Équation 17 : Moyenne de la variable aléatoire X dans un échantillon de taille n.

L'expérience aléatoire qui définit un événement élémentaire m de M consiste à tirer n valeurs de X (de façon indépendante) puis à calculer la moyenne de ces valeurs (Équation 1, page 9).

[1] Cette définition statistique de l'échantillon est généralisable à tous les types de variables aléatoires.

3) Représentations histogrammes de la variable aléatoire M

Une approche par simulation numérique[1] appliquée au cas d'étude 1 permet d'obtenir les histogrammes suivants. Deux paramètres sont pris en compte : le nombre d'échantillons créés au hasard et la taille de ces échantillons (valeur n).

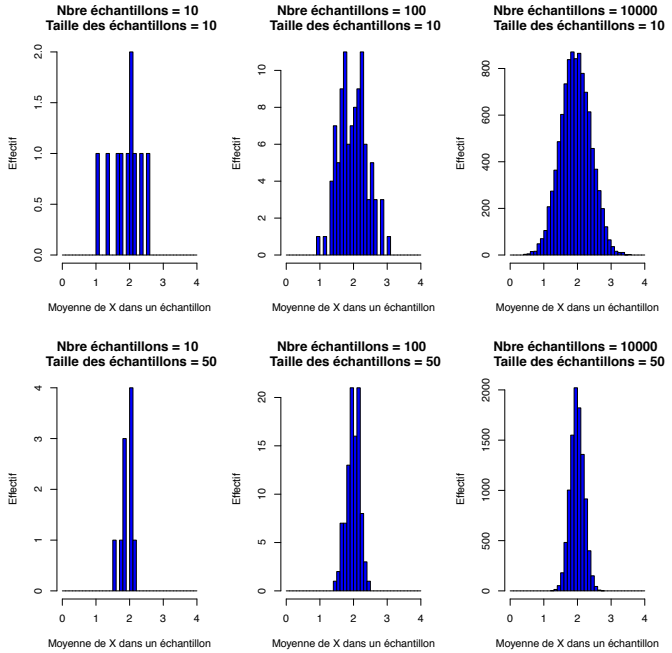

Figure 10 : Résultats de simulations numériques appliquées au cas d'étude 1. La variable aléatoire étudiée est X « écart entre les mesures de la tension artérielle avant et après traitement ». La procédure de création d'un échantillon consiste à tirer n valeurs de X dans l'ensemble (0 ;1 ;2 ;3 ;4). La moyenne associée à ces valeurs est ensuite calculée et représentée en abscisse de ces graphiques. La procédure de création d'un échantillon est répétée plusieurs fois (10, 100 ou 10000 fois). Ainsi, plusieurs ensembles de valeurs moyennes sont obtenus et représentés en histogramme (voir la définition Figure 1, page 10).

➢ **Commenter les résultats des simulations numériques. Quelle est la loi de probabilité de M (VA « moyenne de X dans un échantillon ») ? Quel impact à la taille des échantillons sur cette loi ?**

[1] Ces simulations ont été réalisée avec le logiciel R http://cran.r-project.org/ .

4) Fonction de densité de probabilité associée à la moyenne d'un échantillon

Les résultats des simulations montrent que la loi normale est la fonction de densité de probabilité associée à la variable aléatoire M. La caractérisation précise de cette loi normale nécessite de calculer l'espérance et la variance de M.

(a) Espérance et variance de la moyenne d'un échantillon

➢ **Exprimer l'espérance et la variance de M, en fonction la moyenne μ et de variance σ^2 de X.**

$$E(M) = \mu \quad \text{et} \quad Var(M) = \frac{\sigma^2}{n}$$

Équation 18 : Formules de l'espérance et de la variance de la variable aléatoire « moyenne de X dans un échantillon de taille n » (n étant le nombre d'observations de X). μ et σ^2 sont respectivement la moyenne et la variance de X.

(b) Application du théorème central limite

Les résultats obtenus par la simulation (voir paragraphe précédent) peuvent être « prédits » à l'avance par application du théorème central limite (voir page 21). Ainsi, la moyenne M suit (si n est suffisamment grand) la loi normale suivante :

$$M \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Équation 19 : Loi de probabilité de la variable aléatoire « moyenne de X dans un échantillon de taille n ». Cette formule est valable si n est suffisamment grand (dans la pratique > 30).

Ce résultat rend possible les calculs de probabilités[1] associées à la moyenne d'un échantillon, sans connaitre la loi de probabilité de la variable aléatoire X initialement étudiée.

[1] Ce résultat permet de retrouver toutes les formules utilisées pour la réalisation d'analyses statistiques classiques (estimations et tests d'hypothèses). Il est important de bien le comprendre.

II. Estimation statistique

1) Principe

L'estimation statistique est la procédure d'utilisation des informations obtenues à partir d'un échantillon afin de déduire des résultats concernant l'ensemble de la population.

➢ **Illustrer la notion d'estimation statistique en utilisant le(s) cas d'étude(s). Noter de manière différente les paramètres connus et les paramètres inconnus.**

	Moyenne	Variance	Proportion
Population	μ	σ^2	π
Échantillon	m	Var	p

Tableau 2 : Conventions de notations de paramètres, classiquement utilisés en statistiques à l'Université Paris 7. Les valeurs des paramètres dans la population sont INCONNUES, tandis que les valeurs des paramètres dans les échantillons sont CONNUES (calculés à partir des observations).

2) Qualité d'un estimateur

Un estimateur (tel que la moyenne d'un échantillon) a peu de chances de proposer la valeur exacte du paramètre inconnu. Cela est dû à l'existence d'erreurs d'échantillonnages (une partie de la population a été omise). Pour qu'un estimateur fournisse des estimations utiles, il doit être « sans biais » et « convergent ». C'est le cas de la moyenne d'un échantillon.

(a) Estimateur sans biais

Un estimateur est dit « sans biais » si son espérance est égale à la valeur exacte du paramètre dans la population :

$$E(M) = \mu$$

Équation 20 : Propriété d'un estimateur sans biais, appliquée à la moyenne d'un échantillon. M est l'estimateur et μ le paramètre inconnu de la population.

➢ **Relier cette propriété aux résultats des simulations présentées Figure 10, p. 24.**

(b) Estimateur convergent

Un estimateur est dit « convergent » si sa variance tend vers 0 quand la taille n de l'échantillon augmente :

$$\lim_{n \to +\infty} Var(M) = 0$$

Équation 21 : Propriété d'un estimateur convergent, appliquée à la moyenne d'un échantillon. M est l'estimateur.

Ainsi, les estimations du paramètre μ obtenues à partir de plusieurs échantillons seront d'autant plus proches les unes des autres que la taille n des échantillons est grande.

➢ **Relier cette propriété aux résultats des simulations présentées Figure 10, p. 24.**

(c) Applications à la variance

➢ **La variance d'un échantillon (Équation 2, page 9) est-elle un bon estimateur de la variance σ^2 de X dans la population ?**

3) Estimation ponctuelle

L'estimation ponctuelle consiste à proposer une unique valeur pour le paramètre inconnu. Cette valeur est notée « ^ » :

$$\hat{\mu} = M \qquad \text{et} \qquad \hat{\sigma}^2 = \frac{n}{n-1} Var$$

Équation 22 : Estimations ponctuelles de la moyenne et de la variance d'une variable aléatoire quantitative. Les formules de m et Var sont présentées page 9.

➢ **Les résultats obtenus à partir d'une procédure d'estimation ponctuelle sont-ils fiables ?**

4) Estimation par intervalle de confiance

Construire un intervalle de confiance consiste à donner un intervalle de valeurs dans lequel le paramètre inconnu a une probabilité importante de se trouver. Dans le cas de la moyenne μ, on aura :

$$P(M - e \le \mu \le M + e) = 1 - \alpha$$

Équation 23 : Formule générale de l'intervalle de confiance de la moyenne. μ est le paramètre de la population à estimer et M un estimateur. α est le risque d'erreur associé au calcul de l'intervalle de confiance. Ce risque est dans la pratique fixé à 5%.

➢ **Déterminer la valeur du paramètre e ? (dans le cas où σ^2 est connue[1])**

$$\mu \in \left[M \pm z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}} \right]$$

Équation 24 : Formule de l'intervalle de confiance de la moyenne d'une variable aléatoire quantitative. La valeur de $z_{1-\alpha/2}$ est lue dans la table de la loi normale centrée réduite. Elle dépend de α.

➢ **Quelles sont les valeurs de $z_{1-\alpha/2}$ si $\alpha = 5\%$ (valeur classique), $\alpha = 1\%$ et $\alpha = 10\%$?**

Rappel : $\Phi(z_{1-\alpha/2}) = 1 - \frac{\alpha}{2}$.

III. Cas des petits échantillons ($n \le 30$)

Dans le cas où le nombre d'observations de la variable aléatoire est faible (petit échantillon), le théorème central limite (voir page 21) s'applique relativement mal. Si la variable aléatoire suit une loi normale et que la variance σ^2 est connue, la formule de l'Équation 24 est toujours valable. Cependant, si la variance n'est pas connue, son estimation ponctuelle devra être associée à l'utilisation de la loi de Student. L'intervalle de confiance s'écrit alors :

[1] Dans le cas où la variance n'est pas connue, l'estimateur ponctuel pourra être utilisé. La loi de Student remplacera la loi Normale.

$$\mu \in \left[M \pm t_{\alpha,(n-1)ddl} \sqrt{\frac{\hat{\sigma}^2}{n}} \right] \quad \text{avec} \quad t_{\alpha,(n-1)ddl} \text{ qui est lue dans la table de Student}$$

Équation 25 : Formule de l'intervalle de confiance de la moyenne d'une variable aléatoire quantitative, dans le cas des petits échantillons (ddl étant un nombre de degré de liberté).

IV. Cas des variables aléatoires qualitatives (2 classes)

1) Rappel du second cas d'étude

Constituer un échantillon de taille n consiste (par exemple) à lancer n fois un jeton[1]. La proportion (par exemple de « pile ») est un paramètre qui décrit l'échantillon (comme la moyenne dans le cas d'une variable aléatoire quantitative). On a : $p = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2) Fonction de densité de probabilité de la proportion d'un échantillon

D'après le théorème central limite, la proportion P suit (si n est suffisamment grand) une loi normale :

$$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Équation 26 : Loi de probabilité de la proportion d'une variable aléatoire qualitative à deux classes dans un échantillon. Dans la pratique, il faut pour appliquer ce résultat : $n \times \pi \geq 5$ et $n \times (1-\pi) \geq 5$.

➢ **Quelle relation existe-t-il entre cette formule et la loi de Bernoulli ?**

3) Estimation ponctuelle

L'estimation ponctuelle consiste à proposer une unique valeur pour le paramètre inconnu, ainsi :

$$\hat{\pi} = p$$

Équation 27 : Estimation ponctuelle de la proportion d'une variable aléatoire qualitative à deux classes.

[1] L'unité statistique est ici « un lancer du jeton », la VA est « le côté du jeton obtenu ».

4) Intervalle de confiance

L'intervalle de confiance consiste à donner un intervalle de valeurs dans lequel le paramètre inconnu a une probabilité importante de se trouver, ainsi :

$$\pi \in \left[P \pm z_{1-\alpha/2} \sqrt{\frac{P(1-P)}{n}} \right] \text{ avec } z_{1-\alpha/2} = 1{,}96 \text{ si } \alpha = 5\%$$

Équation 28 : Formule de l'intervalle de confiance de la proportion d'une variable aléatoire qualitative à deux classes.

V. Exemples d'applications

➢ **Calculer les intervalles de confiances associés aux observations présentées ci-dessous en utilisant les 50 observations, puis seulement les 10 premières.**

3 ; 2 ; 0 ; 2 ; 4 ; 3 ; 1 ; 1 ; 3 ; 2 ; 3 ; 1 ; -1 ; 1 ; -1 ; 0 ; 0 ; 4 ; 0 ; -1 ; 3 ; 4 ; 3 ; 2 ; 0 ; 2 ; -1 ; 0 ; 1 ; 1 ; 0 ; 3 ; 0 ; 0 ; -1 ; 0 ; 3 ; 0 ; 0 ; 3 ; 3 ; 4 ; 3 ; 0 ; 2 ; 1 ; 0 ; 2 ; 0 ; 0

Tableau 3 : Ensemble de mesures (ou observations) de la variable aléatoire étudiée dans le cas d'étude 1. Les valeurs écrites en rouge ont été présentées Tableau 1 (page 7). Le nombre d'observations est de 50, la moyenne dans l'échantillon est de 1,3 et la variance de l'échantillon est de 2,37 (soit une variance estimée de 2,42).

Partie 4 : Les tests d'hypothèses

I. Introduction aux tests d'hypothèses

Les tests d'hypothèses reposent sur un raisonnement différent de celui de l'estimation. Il s'agit de faire des hypothèses concernant les valeurs des paramètres d'une variable aléatoire X dans la population et ensuite d'utiliser les observations $(x_1, x_2, x_3, ..., x_n)$ dans un échantillon pour choisir entre les hypothèses.

1) Rappel des cas d'étude

(a) Cas 1 : le médicament est-il efficace ?

- **Définition de la variable aléatoire :** la variable aléatoire X est quantitative (écart entre les mesures de la tension artérielle réalisées avant et après la prise du médicament pendant six semaines).
- **Paramètre de la population étudié :** moyenne μ de la variable aléatoire X dans la population de référence.

- **Hypothèse de travail : le médicament est efficace si la moyenne $\mu > 0$.**

(b) Cas 2 : le jeton est-il équilibré ?

- **Définition de la variable aléatoire :** la variable aléatoire est qualitative à deux classes (côté supérieur – pile ou face – du jeton après avoir été lancé).
- **Paramètre de la population étudiée :** proportion π de côtés « pile » obtenue dans la population de référence.

- **Hypothèse de travail : le jeton est équilibré si la proportion $\pi = \frac{1}{2}$.**

➢ **Discuter les intervalles de confiance calculés au chapitre précédent. Le médicament est-il efficace ? Le jeton est-il équilibré ?**

2) Illustration de l'approche décisionnelle des tests d'hypothèses

Soit X une variable aléatoire quantitative qui suit une loi normale de paramètres inconnus. L'objectif est de choisir entre deux hypothèses nommées H_A et H_B, à partir d'une observation x de X.

Avec : $\begin{cases} H_A : X \sim N(\mu = 0; \sigma^2 = 4) \\ H_B : X \sim N(\mu = 4; \sigma^2 = 16) \end{cases}$

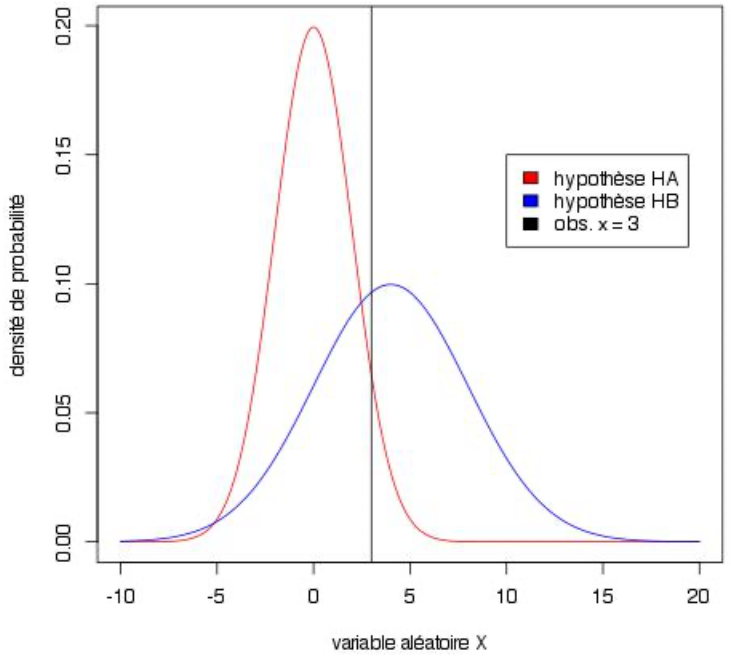

Figure 11 : Illustration de la mise en concurrence de deux hypothèses H_A et H_B. Une observation de X ($x = 3$) est notée par un trait vertical. Le choix entre les hypothèses repose sur une valeur de X observée.

➢ **Quelle hypothèse choisir si l'observation de X est $x = 3$?**

II. Définition d'une règle de décision

1) Objectif

Définir une « règle rationnelle » permettant de choisir entre les deux hypothèses, en fonction d'une observation quelconque x de la variable aléatoire X. Cette règle divisera l'ensemble des valeurs possibles de X en deux sous-ensembles :

- L'ensemble (R_A) des valeurs de X pour lesquelles il faudra choisir l'hypothèse H_A
- L'ensemble (R_B) des valeurs de X pour lesquelles il faudra choisir l'hypothèse H_B

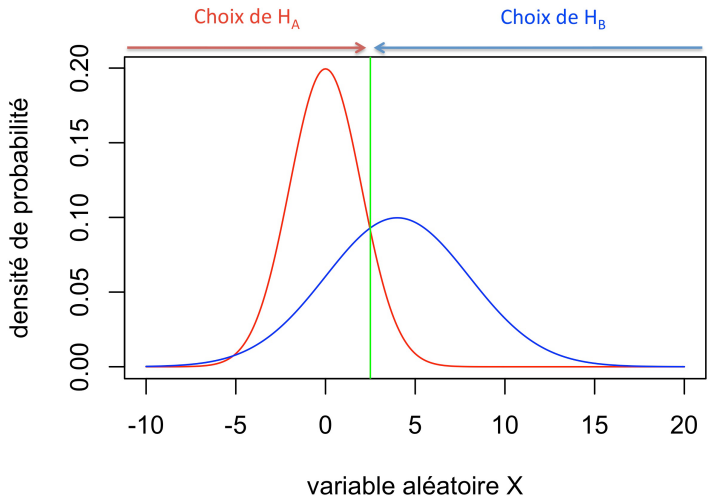

Figure 12 : Exemple de règle de décision applicable au choix des hypothèses H_A et H_B présentées Figure 11. Si $X \leq 2.5$, l'hypothèse H_A est retenue. Si $X > 2.5$ l'hypothèse H_B est retenue. La valeur $x = 2,5$ est à l'intersection entre les deux fonctions de densité de probabilité.

2) Évaluation d'une règle de décision

Évaluer une règle de décision consiste à calculer la probabilité de commettre une erreur sur chacune des hypothèses :

- **Erreur sur l'hypothèse H_A** : choisir H_B alors que H_A est vraie
- **Erreur sur l'hypothèse H_B** : choisir H_A alors que H_B est vraie

➢ **Évaluer la règle de décision : $X \leq 2.5$ choisir H_A et $X > 2.5$ choisir H_B. Deux probabilités sont à calculer : 1) erreur sur l'hypothèse H_A et 2) erreur sur l'hypothèse H_B.**

Il existe de multiples règles de décision. Certaines privilégient l'hypothèse H_A (le risque d'erreur sur H_A est faible), d'autres l'hypothèse H_B (le risque d'erreur sur H_B est faible). Le graphique ci-dessous représente les probabilités d'erreurs associées à une cinquantaine de règles de décision différentes.

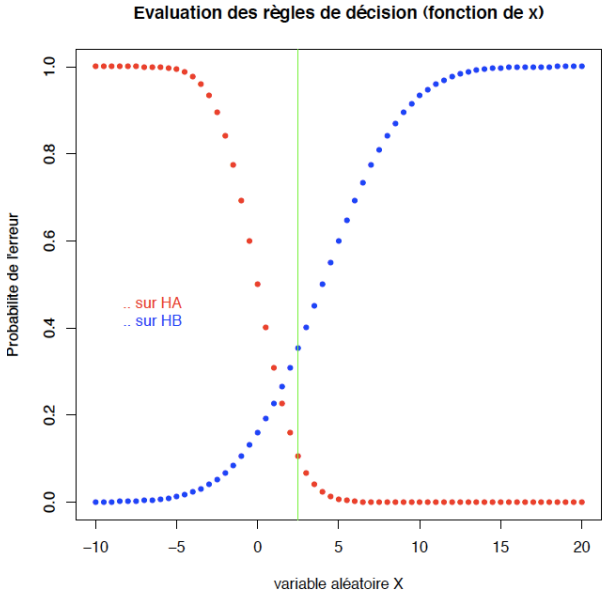

Figure 13 : Calculs des probabilités d'erreurs associées aux hypothèses HA et HB pour différentes règles de décisions. Le trait vertical vert représente la valeur $x = 2.5$, utilisée pour représenter la règle de décision de la Figure 11.

➢ **Discuter la notion de « bonne règle » de décision. Faut-il équilibrer le risque entre les deux hypothèses ? Faut-il privilégier une hypothèse par rapport à l'autre ?**

III. Théorie des tests statistiques

1) Hypothèse nulle et hypothèse alternative

Dans la pratique en statistiques (et en biologie) l'hypothèse selon laquelle seul le hasard est responsable des effets observés dans les échantillons est privilégiée. Cette hypothèse est nommée « **hypothèse nulle** » et notée H_0. L'hypothèse en concurrence est nommée « **hypothèse alternative** » et notée H_1. C'est l'hypothèse selon laquelle le hasard et l'effet biologique testé interviennent.

Hypothèse H_0 (nulle)	Hasard seul
Hypothèse H_1 (alternative)	Hasard + effet testé (efficacité du médicament par exemple)

Tableau 4 : Description des hypothèses mises en concurrence dans la théorie des tests en statistiques.

- **Reprendre les hypothèses nulles et alternatives associées aux deux cas d'études. Discuter les risques d'erreurs associées en termes de faux positifs et de faux négatifs.**

2) Définitions des risques d'erreurs

(a) Erreur sur l'hypothèse H_0

Cette erreur est nommée « risque de première espèce » et notée α. C'est la probabilité de choisir H_1[1] alors que H_0 est vraie.

(b) Erreur sur l'hypothèse H_1

Cette erreur est nommée « risque de deuxième espèce » et notée β. C'est la probabilité de choisir H_0 alors que H_1 est vraie.

[1] Il est de convention en statistiques de dire « rejeter H_0 » plutôt que « choisir H_1 ». Les notions restent toutefois équivalentes.

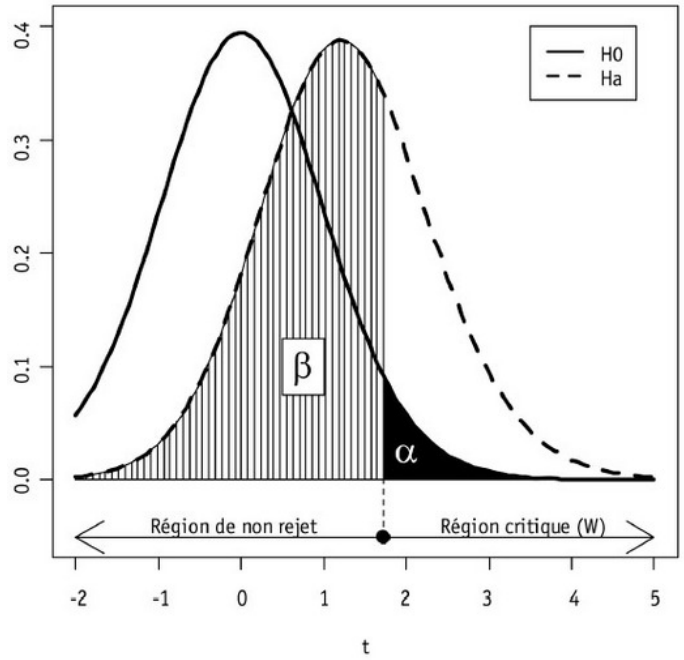

Figure 14 : Représentation des risques statistiques en relation avec la réalisation d'un test d'hypothèses. Le risque α est le risque de première espèce tandis que le risque β est le risque de seconde espèces (les définitions sont données dans le texte principal). Image issue du site Internet : http://www.cairn.info/revue-staps-2007-3-page-49.htm . Ha représente l'hypothèse alternative, notée H_1 dans le cours. La valeur t est la « statistique du test », c'est à dire le paramètre qui est calculé à partir des observations pour choisir entre les hypothèses (voir ci-dessous).

La règle de décision est ensuite choisie de manière à garantir un risque d'erreur α très faible (généralement de 5%). C'est la raison pour laquelle l'hypothèse H_0 est dite « privilégiée » dans la théorie des tests statistiques.

3) Notion de "statistique du test"

Pour choisir entre les hypothèses H_0 et H_1, les paramètres connus de l'étude statistique sont utilisés. Ces paramètres font nécessairement référence aux observations réalisées dans les échantillons. La statistique du test représente la combinaison des paramètres qui va être utilisée pour définir une règle de décision (par exemple "moyenne de X", ou "écart entre les moyennes de X obtenues dans deux échantillons", etc.).

4) Application à la comparaison d'une moyenne et d'une norme (1er cas d'étude)

- **Soit M (moyenne de X dans un échantillon) la statistique du test. Déterminer la limite m_{lim} de M, définissant la règle de décision pour un risque α quelconque.**
- **Les observations présentées dans le Tableau 3 (page 30) sont-elles en faveur d'une efficacité du médicament ? (si $\alpha = 0.05$ et si $\alpha = 0.01$)**

5) Déroulement d'un test d'hypothèses

(a) Formulation des hypothèses et définition de la statistique du test

Les hypothèses d'un test sont formulées en termes de paramètres relatifs à la population étudiée (ces paramètres sont inconnus). La statistique du test est formulée en termes de paramètres relatifs à (ou aux) échantillon(s).

(b) Vérification des conditions d'application

Les conditions d'application d'un test sont en relation directe avec la taille du ou des échantillons utilisés. Elles sont particulières dans le cas des petits échantillons ($n < 30$, voir page 39).

(c) Caractérisation de la fonction de densité de probabilité sous H_0

La loi normale est souvent applicable comme fonction de densité de probabilité. Sous H_0 les paramètres (moyenne et variance) sont généralement connus.

(d) Définition de la règle de décision

La règle de décision dépend d'une part de la statistique du test (paramètre noté en abscisse) et d'autre part du risque d'erreur de première espèce α (positionnement du trait vertical). Deux types de règles de décision s'utilisent, associées aux tests unilatéraux et aux tests bilatéraux (voir ci-dessous).

- **Test unilatéral**

La règle de décision comporte une seule zone de rejet de l'hypothèse H_0 au profit de l'hypothèse H_1.

- **Test bilatéral**

La règle de décision comporte deux zones de rejets de l'hypothèse H_0 au profit de l'hypothèse H_1. Le risque de première espèce α est alors partagé.

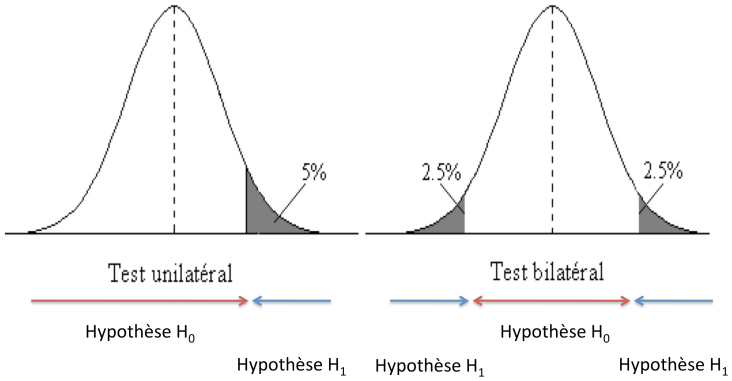

Figure 15 : Représentation des lois de probabilité associées à l'hypothèse nulle H_0 dans le cas d'un test unilatéral (à gauche) et bilatéral (à droite). Sur cet exemple, le risque de première espèce α est fixé dans les deux cas à 5%.

- **Comment changerait la valeur m_{lim} de M (précédemment calculée) avec une règle de décision bilatérale ?**

- **Conclusion**

La conclusion du test repose sur le calcul de la statistique du test à partir des observations de la variable aléatoire dans un échantillon.

IV. Comparaison de deux moyennes

Un nouvel échantillon de 50 patients est constitué. Ces patients ont reçu un second médicament dont l'efficacité doit être comparée au premier. Les nouvelles mesures de la variable aléatoire X sont présentées ci-dessous :

-1 ; 3 ; 3 ; 3 ; 0 ; 3 ; 0 ; -1 ; 3 ; 2 ; 1 ; -1 ; 3 ; 0 ; 1 ; 0 ; 3 ; 3 ; -1 ; 3 ; 1 ; 1 ; 3 ; 0 ; 2 ; 0 ; 0 ; 2 ; 0 ; 1 ; 0 ; 2 ; 4 ; 2 ; -1 ; 1 ; 4 ; 0 ; 3 ; 0 ; 0 ; 4 ; 0 ; 0 ; 2 ; 2 ; 3 ; 2 ; 1 ; 4

Tableau 5: Ensemble de mesures (ou observations) obtenues sur des patients ayant reçu un second médicament. Le nombre d'observation est de 50. Leur moyenne est de 1,4 et l'estimation de la variance de 2,41.

- **Ecrire les hypothèses du test statistique à réaliser. Les discuter en termes d'échantillons et de population(s).**
- **Soit $D = M^{E1} - M^{E2}$ la statistique du test. Déterminer la valeur de l'écart $d_{\lim}$ à partir de laquelle l'hypothèse nulle est rejetée (des effets différents sont associés aux médicaments).**
- **Quelle conclusion tirer à partir des observations présentées dans le Tableau 3 (page 30) et le Tableau 5 (ci-dessus) ?**

V. Cas des petits échantillons

Dans le cas où un des échantillons est composé d'un nombre faible d'observations (généralement inférieur à 30), l'application des tests statistiques présentés ci-dessus nécessite des conditions d'applications particulières.

1) Distribution de la variable aléatoire étudiée

Le théorème central limite s'applique mal si $n < 30$ (voir page 21). Pour que la variable aléatoire « moyenne M d'un échantillon » suive une loi normale (voir page 25), il faut que la variable aléatoire X suive elle-même une loi normale.

2) Égalité des variances

L'estimation de la variance est d'autant plus difficile que la taille du (ou des) échantillon(s) disponibles est petite. Pour éviter de commettre des erreurs, les échantillons sont

généralement rassemblés. Cette étape n'est réalisable que si l'égalité des variances des populations dont sont extraites les échantillons est démontrée[1].

3) Statistique du test

Dans le cas des petits échantillons, la statistique du test à utiliser est présentée ci-dessous (Équation 29). La variance σ^2 est la variance commune, calculée après avoir vérifié l'égalité des variances associées aux deux petits échantillons comparés.

$$T = \frac{M_1 - M_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

Équation 29 : Statistique du test de comparaison de moyennes dans le cas de petits échantillons. Sous l'hypothèse H0, ce paramètre suit une loi de Student à $n_1 + n_2 - 2$ degré de liberté (ddl).

Si l'hypothèse H0 est vraie, le paramètre T suit une loi de Student[2] à ($n_1 + n_2 - 2$ ddl, ci-dessous).

4) Utilisation de la loi de Student

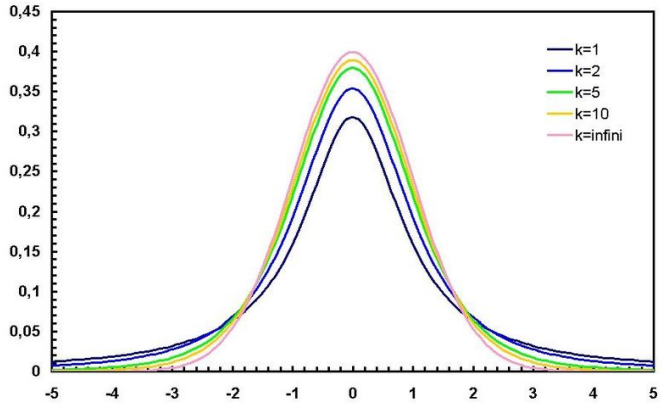

Figure 16 : Fonctions de densité de probabilité de la loi de Student, pour différentes valeurs de degrés de liberté (k). La loi de Student converge vers la loi normale quand k augmente. Cette image est issue du site Internet http://fr.wikipedia.org/wiki/Loi_de_Student .

[1] Cela nécessite la réalisation d'un test statistique : le test de Fisher.
[2] La loi de Student a déjà été évoquée pour le calcul des intervalles de confiance de la moyenne dans le cas des petits échantillons (page 31).

- **Réaliser un test de comparaison des moyennes en utilisant seulement les 10 premières valeurs des échantillons présentés Tableau 3 et Tableau 5. Représenter graphiquement la valeur t_{lim}, lue dans la table de Student et utilisée comme limite pour la règle de décision. Positionner également le risque α.**
- **Les résultats sont-ils différents de ceux obtenus avec l'ensemble des valeurs ?**

Partie 5 : Étude des VA qualitatives multi-classes

I. Exemple de cas d'étude

Une étude statistique supplémentaire est réalisée dans le but d'évaluer le ressenti des patients vis à vis du traitement de leur hypertension artérielle. Les patients doivent indiquer s'ils ont ressenti les effets du traitement de manière « très bénéfique » ; « bénéfique » ; « neutre » ou « mauvais ». Des résultats sont présentés dans le tableau de contingence ci-dessous (Tableau 6).

- **Définir la variable aléatoire de cette étude. Décrire la population et le (ou les) échantillon(s).**
- **Quelle serait les effectifs attendus si la répartition des patients dans les classes suivait une loi uniforme ?**

Effet du traitement	Effectifs observés
Très bénéfique	12
Bénéfique	33
Neutre	52
Mauvais	10
Total	**107**

Tableau 6 : Tableau de contingence associé à l'étude de la variable aléatoire « effet du traitement ressenti ». Cette variable aléatoire est qualitative, à 4 classes. Un échantillon de 107 patients est disponible. Les effectifs observés dans chaque classe sont notés dans ce tableau.

L'analyse des variables aléatoires qualitatives multi-classes repose sur l'étude des tableaux de contingences dans lesquels des effectifs (nombres d'individus) dans chaque classe de la variable aléatoire étudiée.

- **Donner d'autres exemples de variable aléatoire multi-classes.**

II. Test d'hypothèses du χ^2 (Chi2)

1) Principe

Le test d'hypothèse du χ^2 permet d'évaluer si la répartition des individus dans les classes d'une variable aléatoire quantitative multi-classes est compatible avec la répartition attendue selon un modèle théorique (par exemple répartition uniforme).

2) Hypothèses

- H_0 : Le modèle théorique est vrai (c'est l'hypothèse nulle)
- H_1 : Le modèle théorique n'est pas vrai (c'est l'hypothèse alternative)

3) Statistique du test

Le test d'hypothèses du χ^2 utilise comme statistique du test, une mesure d'adéquation (ou critère du χ^2) entre les effectifs observés (notés O_i) et les effectifs théoriques (notés C_i), c'est à dire les effectifs attendus si le modèle théorique est vrai (hypothèse H_0). A noter que pour être utilisé, il faut que toutes les valeurs C_i soient supérieures à 5.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$$

Équation 30 : Mesure d'adéquation utilisée dans le test d'hypothèses du χ^2. Les O_i représentent les effectifs observés et les C_i les effectifs théoriques. k est le nombre de classes de la variable qualitative étudiée.

➢ **Comment est la valeur de la mesure d'adéquation (faible ou forte) si le modèle théorique est vrai ? Pourquoi est-il important de réaliser une division par C_i dans le calcul du χ^2 ? Pourquoi faut-il que les C_i soient > 5 ? Que faire si cela n'est pas le cas ?**

Si l'hypothèse H_0 est vraie, le critère du χ^2 suit une loi du ChI2 (voir ci-dessous Figure 17) pour un nombre de degrés de liberté (ddl) donné :

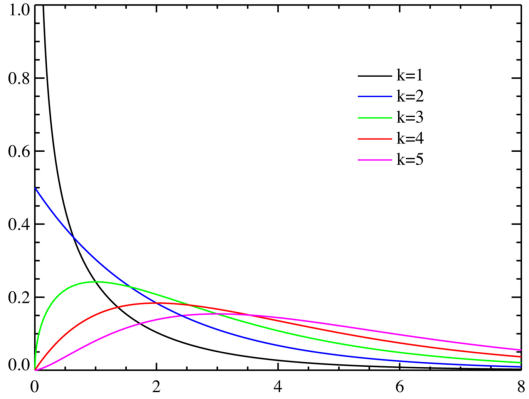

Figure 17 : Fonctions de densité de probabilités du ChI2, pour différentes valeurs de degré de liberté. L'axe des abscisses donne les valeurs de la VA et l'axe des ordonnées les valeurs correspondantes de la densité de probabilité. Cette image est extraite du site Wikipédia https://fr.wikipedia.org/wiki/Loi_du_%CF%87%C2%B2 .

4) Règle de décision

La règle de décision du test dépend alors du risque d'erreur de première espèce α, tel que :

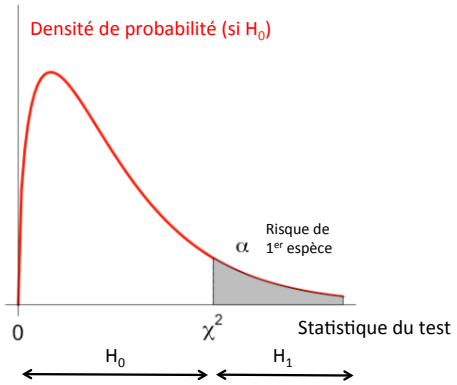

Figure 18 : Règle de décision associée à la réalisation du test du ChI2 (étude d'une VA qualitative multi-classe). La valeur limite peut être lue dans une table statistique (pour risque et un nombre de ddl donné).

➢ **Appliquer le test du χ^2 aux observations présentées dans le Tableau 6 ; pour un modèle théorique de loi uniforme. Commenter les résultats.**

III. Pour aller plus loin

Il existe une variante du test de χ^2 (ci-dessus) et qui consiste à comparer les répartitions d'individus dans les classes de la variable entre deux échantillons. Le modèle théorique consiste à considérer que les répartitions sont les mêmes.

Une illustration de ce test pourrait être de considérer de manière séparée, les hommes et les femmes qui ont participé à la précédente étude (voir ci-dessus, Tableau 6). La question serait alors "y a-t-il un ressenti des effets du médicament, différent entre les hommes et les femmes ?".

Effet du traitement	Femmes	Hommes	Total
Très bénéfique	9	3	12
Bénéfique	11	22	33
Neutre	10	42	52
Mauvais	9	1	10
Total	**39**	**68**	**107**

Tableau 7 : Tableau de contingence associé à l'étude de la variable aléatoire « jour de la naissance ». Cette variable aléatoire est qualitative, à 7 classes. Deux échantillons de 4231 individus et 4038 individus sont disponibles. Il s'agit cette fois de comparer la répartition des naissances entre les années 1 et 2.

➢ **Appliquer le test du χ^2 aux observations présentées ci-dessus. Définir le modèle théorique, calculer les effectifs attendus (selon ce modèle), la statistique du test et conclure.**

Partie 6 : Évaluer les performances d'une analyse statistiques

I. Précision d'un intervalle de confiance

L'intervalle de confiance permet d'évaluer la précision de l'estimation d'un paramètre statistique d'une population, à partir des observations d'un échantillon. Pour un risque d'erreur donné (par exemple $\alpha = 5\%$), plus l'intervalle de confiance est resserré autour de l'estimation ponctuelle du paramètre d'intérêt plus il est « précis ».

Dans le cadre d'une étude statistique, la précision d'un intervalle de confiance peut être utilisé de deux manières :

- Compte tenu d'une taille d'échantillon donnée, quel est le taux de précision de l'estimation obtenue ?
- Compte tenu d'un taux de précision souhaité, quelle est la taille d'échantillon nécessaire pour l'obtenir ?

➢ **Quelle doit être la taille de l'échantillon pour réduire de moitié l'amplitude des intervalles de confiance calculés à partir des données du Tableau 3 (page 53).**

II. Degré de signification d'un test statistique

Lorsqu'un test statistique aboutit au rejet de l'hypothèse H_0, le degré de signification (nommé « p-value » en anglais) mesure la confiance qu'il est possible d'avoir en cette conclusion, par rapport à l'arbitraire de $\alpha = 5\%$. C'est la valeur limite ($< 5\%$) qui aurait pu être prise comme risque d'erreur sur H_0, tout en conservant la conclusion de rejet de l'hypothèse H_0.

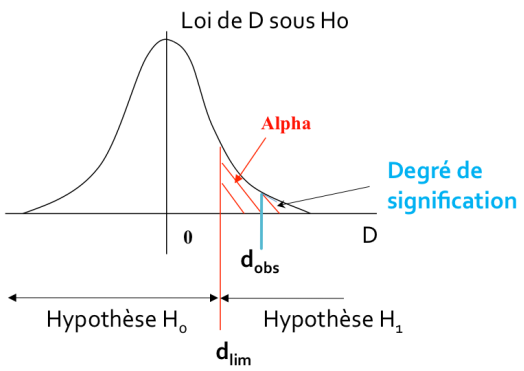

Figure 19 : Illustration du degré de signification associé à un test statistique. Le degré de signification est noté p (*p-value* en anglais). Plus sa valeur est faible, plus le test est dit « significatif ».

Ainsi, le calcul de la pvalue (dans le cas d'un test unilatéral) est :

$$pVal = \Pr(D > d_{obs})$$

Équation 31 : Formule utilisée pour calculer le degré de signification d'un test statistique. Ici D est le paramètre utilisé pour la statistique.

- **Calculer les degrés de signification associés aux tests statistiques précédemment réalisés. Discuter la notion de règle de décision, en quoi le choix d'une valeur fixe de risque α est-il arbitraire ?**

III. Notion de puissance d'un test statistique

La puissance d'un test statistique est la probabilité de rejeter H_0 alors que H_1 est vraie (détecter un effet – par exemple efficacité d'un médicament ou truquage d'un jeton – alors que cet effet existe réellement). La puissance est directement reliée au risque de 2nd espèce β : $P = 1 - \beta$.

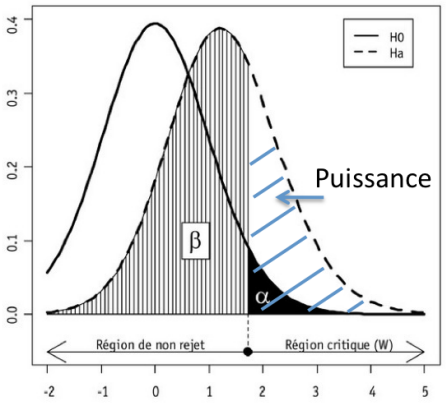

Figure 20 : Illustration de la puissance associée à la règle de décision d'un test statistique. La puissance est la probabilité de rejeter H_0 alors H_1 est vraie, elle est donc associée à la loi sous H_1.

- **Calculer les puissances des tests statistiques réalisés à partir des données du Tableau 3 et du Tableau 8 (page 53).**

IV. Simulations appliquées au cas d'étude 2 (logiciel R)

1) Combien de fois faut-il lancer un jeton ?

- **Discuter les graphiques présentés en annexes.**

2) Combien de jetons faut-il lancer ?

- **Discuter les graphiques présentés en annexes.**